Report: Agent Selection Algorithm

Introduction

Agent selection algorithm aims to match user query with ‘best’ agent.

Goal:

- Input: (query, set of agent descriptions), and (response time, input cost, output cost, average rating, rated responses, popularity)
- Output: 'Best' agent

Feature engineering:

- **Quality score:** average rating provide insight to how users feel on ave, it is unreliable when rated_responses is low. So we scale average_rating with rated_responses and call it quality_score:

$$\text{Quality Score} = \frac{\text{average\_rating} \times \text{rated\_responses}}{\text{rated\_responses} + k}$$

  where smaller k values make the quality score converge to the average rating faster with respect to rated responses. When k = 0, the quality score equals the average rating.

- **Adjusted quality score:** Quality score is still flawed. If there agent with 0 rated_reponse, quality_score become zero giving agent not chance to be selected. We would want those agent some chance and come up with adjusted_quality_score:

$$adjusted\_quality\_score = \frac{\text{average\_rating} \times \text{rated\_responses} + \text{baseline\_rating} \times k}{\text{rated\_responses} + k}$$

  This still converge to average_rating and give newer agents a chance. However, this metric dilute the impact of well-rated agent. If we set the `baseline_rating` to 5, any agent with zero rated responses would receive a `quality_score` of 5, making them indistinguishable from agents who consistently earn 5-star ratings from users. So lower `baseline_rating` is recommended.

- **log popularity:** The impact of popularity should be more significant for a change from 0 to 100 users than from 10,000 to 10,100 users. So we scale them by log:

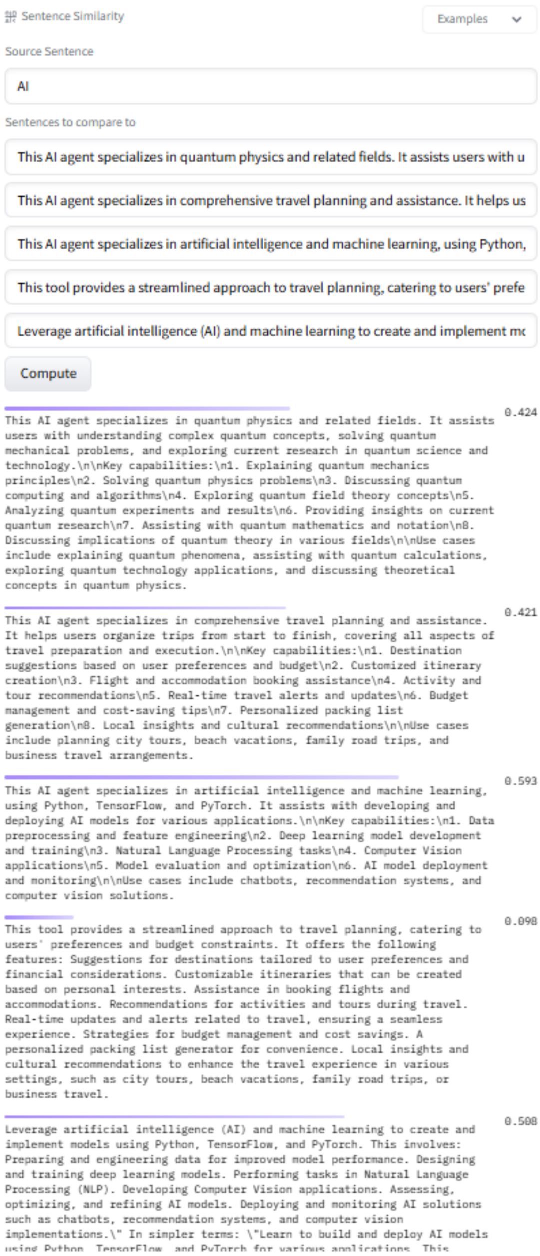
$$log\_popularity\_score = log(popularity + 1)$$

- **total estimated cost:** Ideally, we'd use average input and output tokens to scale the total estimated cost, but we currently don’t have this data.

$$Total\_estimated\_cost = input\_cost + output\_cost.$$

- **Processed description:** Instead of using the given description directly we
  1. remove some phrase that may cause bias getting cleaned_description
  2. then rephrase the cleaned_description using LLM.

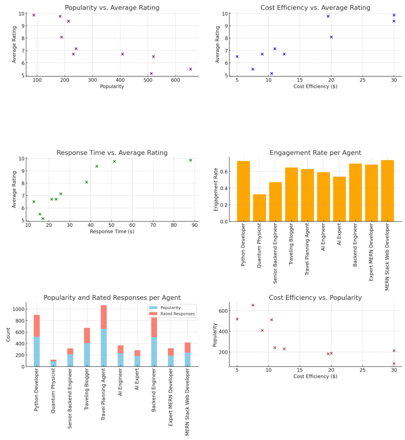
Why? Without processing, phrases like “This AI agent” may cause irrelevant matches. For example below, a user query “AI” could match both a quantum physicist and a travel agent simply due to the generic “AI” mention. Processing eliminates these false positives, ensuring more accurate matching.



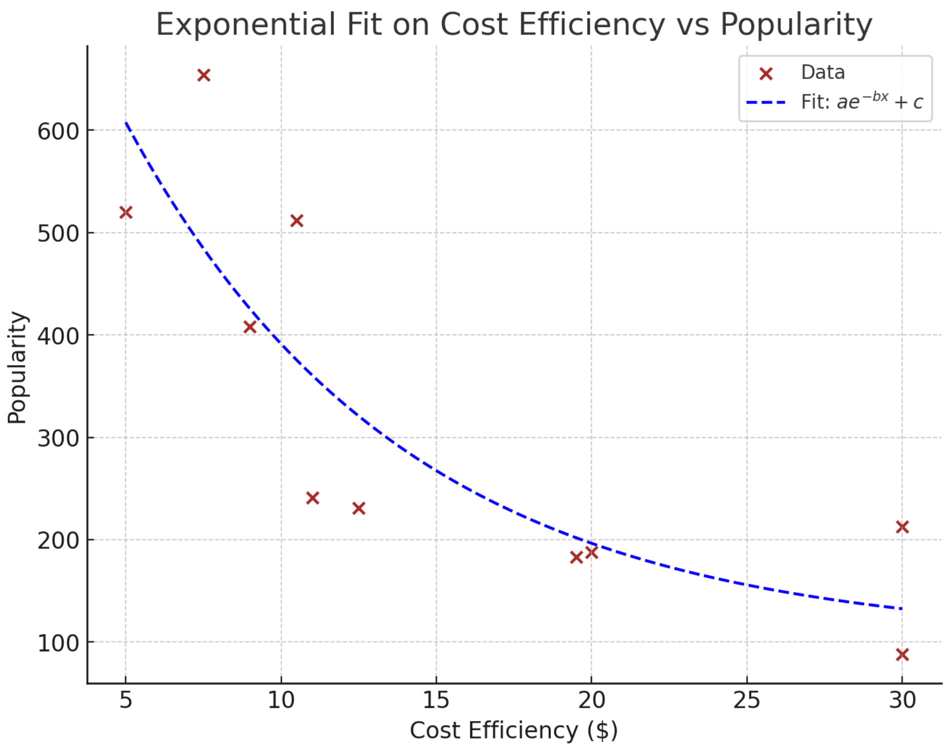
Feature analysis (on non-benchmark data)

This is done on data in `data/agents` folder

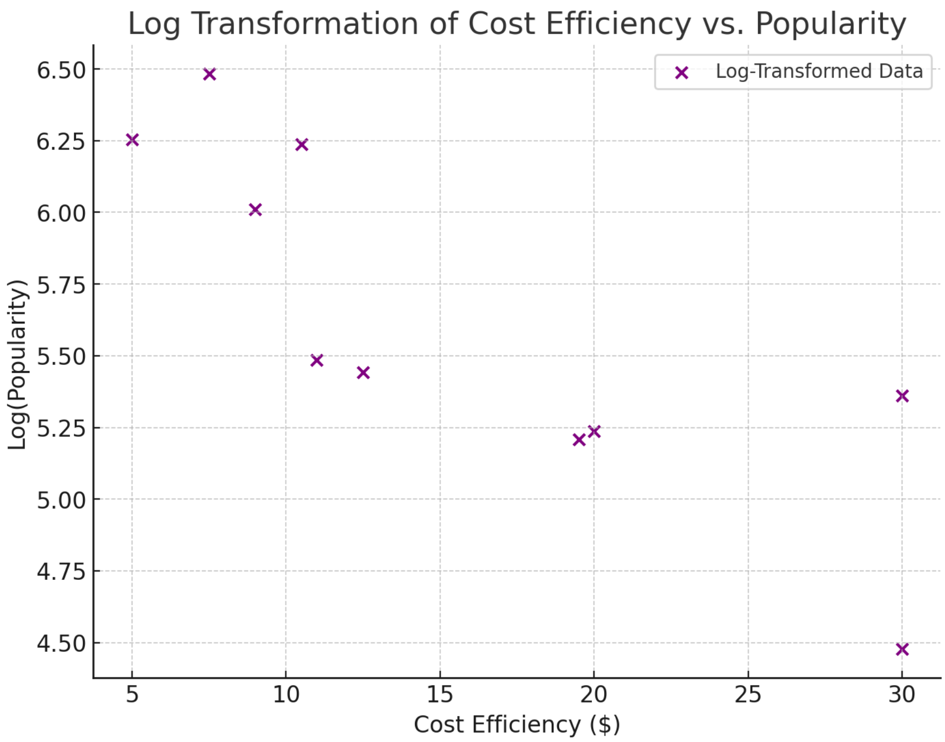
There seem to be linear relationship between many of these features. We are not sure how we should utilize them though.(cost efficiency is total estimated cost.... I know it was a bad naming)



1. Agent with low average rating seem to have high popularity. Very counter intuitive.
2. The more expensive agent seem to have higher average rating. make sense.
3. Agent with higher response time seem to have high average rating.
4. From 2 and 3, There must be relationship between total estimated cost and repsonse time.



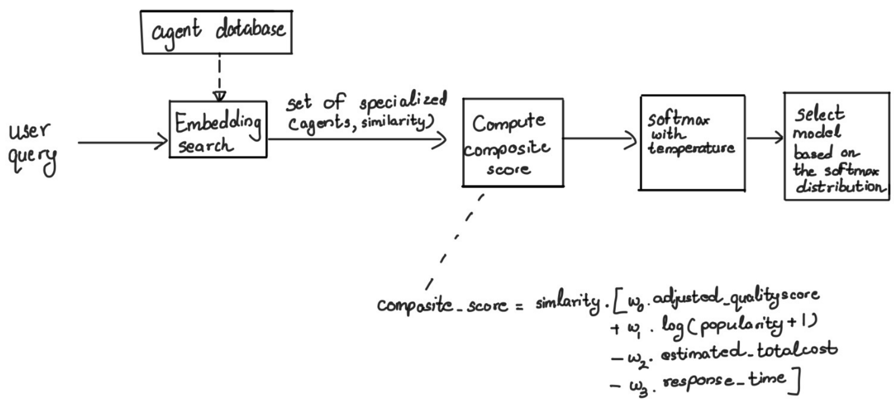
- Users seem to query agent with lower cost more, exponentially. This support our decision to use log scale on popularity.



- log_popularity seem to result in linear relationship.

Agent selection

After discussing features, we should look into our main algorithm.



- Embedding search
  - **Embedding function:** We use sentence-transformer text embedding (all-mpnet-base-v2) with TFIDF.
    - Neural network for *semantic search*. TFIDF for *lexical search* (with snowflake stem preprocessing). lexical search is particularly suitable for this task, since user query is (probably) not very long.
    - We also try SPLADE which give good result: SPLADE for Sparse Vector Search Explained | Pinecone
    - We precompute embedding of agent description for later use when user query.
  - **Similarity function:** We use cosine similarity (why? Pretrained Models — Sentence Transformers documentation)

    - **Efficient embedding search**: We use **Hierarchical Navigable Small Worlds (HNSW)** algorithm. Which trade setup time and high memory usage for fast search, add, remove time.
    - We also try some other search algorithm using faiss (Faiss: The Missing Manual | Pinecone). However, we think HNSW is ok for now in term of both efficiency and accuracy.
- Composite score

  While similarity between query and agent is important, but when many agents have similar description, additional metrics can help decide between them.

    - We combine `adjusted_quality_score`, `log_popularity`, `estimated_total_cost`, and `response_time` in a weighted sum to calculate the composite score. These weights are currently intuition-based.
    - Ideally, these weights would be optimized based on user preferences or measurable objectives. If relationships are complex, a machine learning model could replace the composite score.
    - Then, we have 2 possible approaches:
        1. **Fixed Agent Pool:** Use a fixed number of agents and multiply `composite_score` by `similarity_score` to obtain an `adjusted_composite_score`. This approach prioritizes query-description similarity while still factoring in other metrics.
        2. **Similarity-Filtered Pool:** Filter agents based on a threshold relative to the maximum similarity score, ensuring only agents highly similar to the query are considered.

        ```
        max_similarity = max(agent["Similarity Score"] for agent in scored_agents)
        filtered_agents = [agent for agent in scored_agents if agent["Similarity Score"]
        > 0.8 * max_similarity]
        ```

Selection:

When multiple agents have exactly same descriptions, one may have higher ratings or popularity than others. We would want to prioritize agents with higher scores, while allowing others a chance as well.

- A straightforward approach is to select the agent with the highest `composite_score`. However, this may exclude agents with similar capabilities but slightly lower ratings or popularity.
- Alternatively, applying a softmax function to the `composite_score` generates a probability distribution, enabling us to sample agents and give those with slightly lower scores a chance of selection.

- Why not cross-encoder?
    - We tried but results are not as good as embedding in our case. It does improve some cases though as we will see in results.
    - Most cross-encoders are optimized for sentence pairs with exact meaning matches, which is different from our goal of matching user questions with agent descriptions.
        - For reference, many cross-encoder models are trained on datasets like SNLI and MultiNLI, where sentence similarity is key.

Result

In the benchmark, we selected the agent with the highest similarity score. For models using processed descriptions, we ran 5 trials to account for output variability introduced by the LLM.

We used the following models:

- **Embedding Model 1**: Ensemble of all-mpnet-base-v2 with TFIDF, ngram_range=(2,3)
- **Embedding Model 2**: Ensemble of Transformer + SPLADE
- **Cross-Encoder Model**: ms-marco-MiniLM-L-6-v2

Approach	Run 1	Run 2	Run 3	Run 4	Run 5	On average

Embedding Model 1	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
Cross Encoder	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833
Embedding Model 1 with processed description	0.6250	0.7083	0.6667	0.6667	0.6250	0.65834
Cross Encoder with processed description	0.5833	0.5833	0.5417	0.5833	0.6250	0.5833333333333334
Embedding Model 2	0.5000	0.5000	0.5000	0.5000	0.5000	0.5
Embedding Model 2 with Processed Description	0.7083	0.7083	0.7083	0.7500	0.6250	0.7

Analysis

Embedding vs. Cross-Encoder:

- The basic embedding model scores 0.5 consistently, while the cross-encoder improves slightly with an average score of 0.5833.
- However, the cross-encoder scores lower than the embedding model when descriptions are processed, suggesting it may not capture relevance as effectively under these conditions.

Effect of Processed Descriptions:

- Processing descriptions significantly improves the embedding model's performance, reaching an average score of 0.6583, which indicates that processing enhances relevance by removing irrelevant phrases.
- The **highest score with processed descriptions** (0.7083) suggests potential for further improvements in the processing pipeline.

SPLADE Performance:

- The SPLADE-based embedding with processed descriptions achieved the highest overall average (0.7) and consistently outperformed other models in every run, highlighting its ability to handle complex descriptions and match query relevance effectively.

Overall Performance

- The **SPLADE-based embedding with processed descriptions** yields the best performance (average of 0.7 and a peak of 0.7500), indicating it as the most effective approach.
- This method does require additional setup time for processing descriptions when adding new agents, but it compensates with better accuracy and relevance in matching.